ChinaXiv合作期刊

基于案例推理和启发式 Q 学习的资源分配算法 *

徐　琳，赵知劲

(杭州电子科技大学通信工程学院，杭州 310018)

摘　要：针对集中式认知网络的信道和功率分配问题，提出了一种基于案例推理和启发式Q学习算法。为了提高Q学习算法的收敛速度，将当前分配问题与存储的历史案例进行相似度匹配，选取最相似案例的Q值，归一化处理后作为启发式Q学习算法的初值。为了提高启发式Q学习的算法性能，引入一个基于信息强度的指导函数，通过强调动作的重要性来改变动作策略；设计的奖赏函数反映了认知系统的能量效率。仿真结果表明，该算法可以明显提高认知网络信道和功率分配的认知系统能量效率和收敛速度。

关键词：信道和功率分配；启发式Q学习；案例推理；认知无线电；认知系统能量效率；成功传输概率

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.07.0416

Resource allocation algorithm based on case reasoning and heuristically accelerated Q-learning

Xu Lin, Zhao Zhijin

(*Telecommunication School*, *Hangzhou Dianzi University*, *Hangzhou* 310018, *China*)

Abstract: Aiming at the problem of channel and power allocation in centralized cognitive networks, a case-based reasoning and improved heuristically accelerated Q-learning algorithm was proposed. In order to improve the convergence speed of the Q learning algorithm, the current allocation problem was matched with the stored historical case, and the Q value of the most similar case was selected, which was normalized as the initial value of the heuristically accelerated Q learning algorithm. In order to improve the performance of the heuristically accelerated Q learning, a guidance function based on information intensity was introduced to change the action strategy by emphasizing the importance of the action; energy efficiency was considered in the design of the reward function. The simulation results show that the proposed algorithm can significantly improve the system energy efficiency and convergence speed, which has carried the channel and power allocation.

Key words: channel and power allocation; improved heuristically accelerated Q-learning; case reasoning; cognitive radio; system energy efficiency; successful transmission probability

chinaXiv:201810.00020v1

0　引言

无线电频谱是无线通信中最宝贵的资源，固定的频谱分配政策导致频谱资源紧张，同时又存在大量频谱处于空闲状态[1]。认知无线电技术的提出大幅度提高了频谱利用率，它允许认知用户机会式利用空闲信道，通过与周围环境的交互，对信道、发射功率等参数进行动态分配和优化，以满足更多用户的通信需求[2]。

强化学习算法以环境状态为输入，奖惩信号为反馈，通过agent与环境不断交互学习，从而输出最佳的决策策略[3]。Q学习是使用最广泛的强化学习算法，已经成功应用于集中式网络架构的认知无线电资源分配问题中。Yao 等人[4]首次利用经典Q学习（Q-learning, QL）完成了集中式网络架构下信道和功率的联合分配，但收敛速度较慢。伍春等人[5]对信道和功率进行分层处理，各认知用户根据信道增益选择信道后，采用可变学习速率Q学习（Q learning with Variable learning rate, VLRQL）进行功率分配，应用该算法的系统容量略低，收敛速度有待提高。针对这些问题，本文提出了一种基于改进启发式 Q 学习（improved heuristically accelerated Q-learning，IHAQL）算法。在启发函数加快Q学习的基础上，引入基于信息强度的指导函数，根据奖赏值更新信息强度，从而改变动作策略，提高收敛速率。应用案例推理（case-based reasoning，CBR）能够明显提高认知无线电NC-OFDM（non-contiguous OFDM）的资源分配的收敛速度[6,7]。因此，为了优化合作Q学习算法的Q值初始化，进一步提高算法性能，本文提出了基于案例推理和改进启发式 Q 学习（case-based reasoning and improved heuristically accelerated Q-learning, CBR-IHAQL）的集中式认知网络的信道和功率分配算法。利用案例推理，选择与当前问题最相似案例

收稿日期：2018-07-25；**修回日期：**2018-09-11　**基金项目：**国防"十二五"预研项目

作者简介：徐琳（1994-），女，浙江台州人，硕士研究生，主要研究方向为认知无线电、信号处理（273678008@qq.com）；赵知劲（1959-），女，浙江宁波人，教授，博导，博士，主要研究方向为认知无线电、通信信号处理、自适应信号处理等.

ChinaXiv合作期刊

的Q值，作为后续改进Q学习的初始值，使得各Agent在学习初就接近最优解。本算法明显提高了收敛速度，且能获得更高的认知系统能量效率。

1　案例推理和改进启发式Q学习算法

1.1　CBR-IHAQL算法模型

当有新案例到达，先与案例库中的历史案例进行匹配，选取最相似案例的Q值归一化作为Q学习的初始值，然后进行改进启发式Q学习。新案例得到解决后，将新案例的效用值与案例库中效用值最小的案例进行比较，从而决定是否进行案例更新。CBR-IHAQL算法模型如图1所示，虚线框中为案例推理算法。

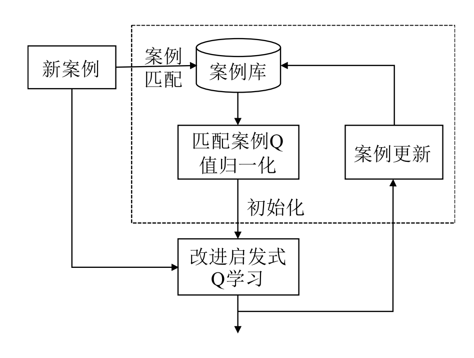


图1 CBR-IHAQL算法模型

1.2　案例推理

案例推理可以为决策过程提供决策知识和智能信息服务，其技术优势是可以累积专业知识，为问题的识别和解决提供适当的建议[8]。典型的案例推理包括案例分析和表示、案例匹配、案例修正和案例更新四个部分，其过程一般为通过新问题和历史案例的相似度评估检索最佳案例，修正后作为新问题的解决方案，如图2所示[9]。

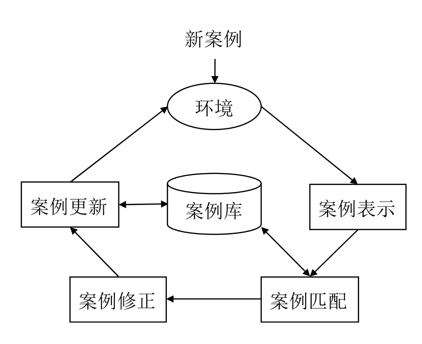


图2　案例推理算法流程

假设案例库中的案例由图3所示四部分组成。

ID	特征参数向量*X*	解决方案*Y*	效用值*E*

图3　案例的存储结构

其中，ID为案例在案例库中的存储标号；解决方案和效用值E通常根据应用需求来设定；特征参数用向量*X*表示，案例的特征向量具体表示如式（1）所示，

$$X_k=(x_k^1,x_k^2,\ldots,x_k^D),\quad k=1,2,\ldots,L \tag{1}$$

其中：D为案例的特征总数；L为案例库中案例个数。利用欧式距离作为案例的匹配函数。

假设新案例为c_{new}，则新案例与历史案例c_k的相似值为

$$f(c_{new},c_k)=\sum\nolimits_n \delta_n \sqrt{\sum\nolimits_j (x_{new}^{n,j}-x_k^{n,j})^2} \tag{2}$$

其中：δ_n为第n个特征参数的权值；$x_{new}^{n,j}$和$x_k^{n,j}$分别为c_{new}和c_k的第n个特征的第j个参数值。

因此可得，匹配案例为

$$\arg\max_{c_k} f(c_{new},c_k) \tag{3}$$

1.3　传统启发式Q学习

Q学习是一种最常用的在线学习技术[10]，通过Agent的不断试错与环境进行交互，利用在交互过程中产生的奖赏来改进学习策略，进而寻得最佳策略[11]。Q学习通常以状态-动作值函数$Q_t(s,a)$作为评估动作优劣的依据，状态s_t下，选择动作a_t时，其Q值更新如式（4）所示。

$$Q_{t+1}(s_t,a_t)\leftarrow(1-\alpha)Q_t(s_t,a_t)+\alpha(r_t+\gamma\max_a Q_t(s_{t+1},a)) \tag{4}$$

其中：$0<\alpha<1$为学习速率；$0<\gamma<1$为奖赏折扣值，表示下一状态对当前状态下Q值的影响程度；r_t为执行选择的动作所获得的奖赏值；s_{t+1}为下一状态；a_{t+1}为下一状态中Q值最大的动作。

为了提高Q学习算法的收敛速度，结合启发函数进行动作的选取，以更有效的方式来指导agent对状态动作空间的探索。通常，启发函数只作用于动作策略，其最大特点是函数值$H(s,a)$不断地进行在线更新，以此来突出表现优秀的动作。传统启发函数$H(s_t,a)$的更新式如下[12]，

$$H(s_t,a)=\begin{cases}\max\limits_{a'} Q(s_t,a')-Q(s_t,a)+\theta & a=\pi^H(s_t)\\ 0 & 其他a\end{cases} \tag{5}$$

其中：θ是一个较小正实数，$\pi^H(s_t)$是启发函数H建议的最佳动作。

1.4　改进启发式Q学习

1.4.1 指导函数设计

为了进一步加快Q学习的收敛速率，减少对不必要动作的探索，在动作策略中引入一种基于信息强度的指导函数$G(s,a)$，其设计结合了Q函数和H函数，利用信息强度对动作的重要性进行评估[13]，在启发式函数的基础上对动作的选择进行进一步的指导，定义式如下所示，

$$G(s_t,a)=\begin{cases}\max\limits_{a'}(Q(s_t,a')+H(s_t,a'))-(Q(s_t,a)+H(s_t,a)) \\ \quad +U\dfrac{p(s_t,a_t)}{\sum_k p(s_t,a_k)} & a_t=\pi^p(s_t)\\ 0 & 其他\end{cases} \tag{6}$$

其中：$p(s_t,a)$为状态s_t下动作a的信息强度；$\pi^p(s_t)$是信息强度函数p建议的最佳动作；$\dfrac{p(s_t,a_t)}{\sum_k p(s_t,a_k)}$反映当前动作的重要性；$U$表示信息强度对动作策略影响度量。

chinaXiv:201810.00020v1

ChinaXiv合作期刊

chinaXiv:201810.00020v1

信息强度 $p(s_t,a)$ 的更新式如式（7）所示。

$$p(s_t,a)=\begin{cases} p(s_t,a)\dfrac{r_{\max}}{r_t} & a\neq a_t \\ 1 & 其他 \end{cases} \tag{7}$$

其中：$r_{\max}$ 是状态 s_t 下之前记录的最大奖赏值。

信息强度 $p(s_t,a)$ 的更新是由 $r_{\max}$ 和 r_t 的大小决定的，当 $r_t > r_{\max}$ 时，表示当前选择的动作比之前记录的最佳动作更优，因此要对该状态下所有动作的信息强度按式（7）进行更新，否则无需更新。以上信息强度的更新规则表明，在保留之前信息强度的同时，根据 $r_{\max}$ 和 r_t 的大小更新的信息强度可以体现出各动作的优劣性。

1.4.2 动作选择策略

Q 学习算法的动作策略选择时，通常要考虑权衡搜索和利用。若侧重搜索，会增加找到最优解的概率，但算法的收敛速度会较慢；若侧重利用，会加快算法的收敛，但容易陷入局部最优。最常用的是动作选择策略是 ε -贪婪策略和 Boltzmann 机制。本文在 Boltzmann 机制中引入启发函数和指导函数，提出一种改进的动作策略，其更新式如式（8）所示。

$$\pi(s)=\arg\max_{a_i}\frac{e^{[Q(s_t,a_i)+H(s_t,a_i)+G(s_t,a_i)]/T}}{e^{\sum_k [Q(s_t,a_k)+H(s_t,a_k)+G(s_t,a_k)]/T}} \tag{8}$$

其中：*T*>*0* 为温度参数。*T* 较大时所有动作都能被等概率地选取；随着 *T* 的减少，将以最大概率选取 Q 值与 H 值之和最大的动作。

1.5 改进启发式 Q 学习算法的收敛性分析

本节证明上述改进启发式 Q 学习算法的收敛性。

证明　假设在状态 s^*，其记录的最优动作为 a_1，在学习过程中选择动作 a_2 获得了更大的奖赏值，则根据式（7）可得信息强度 $p(s^*,a_1)<p(s^*,a_2)$，则

$$\pi^p(s^*)=\max_a\max_a p(s^*,a)=a_2$$

a）当 $a=a_2$ 时，根据式（5）启发函数和式（6）指导函数的更新规则可得，

$$H(s^*,a_2)=\max_a\max_a Q(s^*,a)-Q(s^*,a_2)+\eta \tag{9}$$

$$\begin{aligned}G(s^*,a_2)=&\max_a\left(Q(s^*,a)+H(s^*,a)\right)-\\&[Q(s^*,a_2)+H(s^*,a_2)]+U\frac{p(s^*,a_2)}{\sum_k p(s^*,a_k)}\end{aligned} \tag{10}$$

b）当 $a=a^*$ 时，其中 a^* 为包含 a_1 但不包含 a_2 在内的动作，

$$H(s^*,a^*)=0 \tag{11}$$

$$G(s^*,a^*)=0 \tag{12}$$

利用式（9）（10）可得，

$$\begin{aligned}&Q(s^*,a_2)+H(s^*,a_2)+G(s^*,a_2)\\&=Q(s^*,a_2)+H(s^*,a_2)+\max_a\left(Q(s^*,a)+H(s^*,a)\right)\\&\quad-[Q(s^*,a_2)+H(s^*,a_2)]+U\frac{p(s^*,a_2)}{\sum_k p(s^*,a_k)}\\&=\max_a\max_a\left(Q(s^*,a)+H(s^*,a)\right)+U\frac{p(s^*,a_2)}{\sum_k p(s^*,a_k)}\end{aligned} \tag{13}$$

其中 $U\dfrac{p(s^*,a_2)}{\sum_k p(s^*,a_k)}>0$，$H(s^*,a)\geq 0$。

利用式（11）（12）可得，

$$Q(s^*,a^*)+H(s^*,a^*)+G(s^*,a^*)=Q(s^*,a^*) \tag{14}$$

比较式（13）（14）可得

$$\begin{aligned}&Q(s^*,a_2)+H(s^*,a_2)+G(s^*,a_2)\\&>Q(s^*,a^*)+H(s^*,a^*)+G(s^*,a^*)\end{aligned} \tag{15}$$

因　为　$\max_{a_i}\dfrac{e^{[Q(s_t,a_i)+H(s_t,a_i)+G(s_t,a_i)]/T}}{\sum_k e^{[Q(s_t,a_k)+H(s_t,a_k)+G(s_t,a_k)]/T}}$　等　价　于 $\max_{a_i}[Q(s_t,a_i)+H(s_t,a_i)+G(s_t,a_i)]$，则由式（8）（15）可知，在利用阶段

$$\pi(s^*)=a_2$$

由上述证明可知，该算法的动作策略收敛于信息强度大的策略，且通过不断学习更新，必将收敛于最优策略。信息强度的更新，可以指导 Agent 选取更优秀的动作，减少不必要的探索，从而进行更有效的学习。

2　认知无线电资源分配算法

2.1　系统模型

本文将案例推理和改进启发式 Q 学习算法用于集中式认知无线电网络的信道和功率分配问题。其中，集中式网络结构如图 4 所示。假设网络中存在 M 个主用户（PU），K 个认知用户（SU），以及 N 个可用于主用户和认知用户的信道。主用户以概率 λ 在其信道上传输信息，各信道只能由一个主用户或认知用户占用。各认知用户能准确感知主用户的通信，并反馈给中心基站。

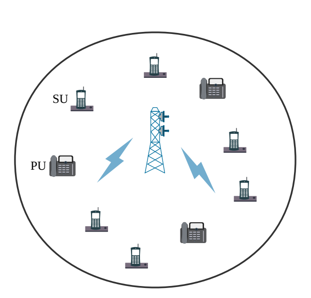


图 4　集中式网络拓扑结构

ChinaXiv合作期刊
chinaXiv:201810.00020v1

2.2　状态-动作空间和奖赏函数的设计

本文将中心基站被视为学习 Agent，定义状态 $s_t = cr_i$，$i \in \{1,2,...,K\}$，cr_i 是认知用户的编号。通过一次次迭代对状态的遍历，Agent 可以为每个认知用户分配信道和功率，并且逐步优化。

动作的设置考虑信道和功率的联合分配，即 $a_i = [channel, power]$，$channel \in \{1,2,...,N\}$ 为可选信道，$power \in \{P_1, P_2, ..., P_z\}$ 为可选传输功率，$P_1 < P_2 < \cdots < P_z$，z 为功率的总类数。

本文算法的目标是用户能成功通信，且系统能量效率尽可能最大化，因此奖赏函数 r_t 定义如下，

$$r_t = \begin{cases} -5 & \text{产生冲突} \\ 1 + \dfrac{W \log_2(1 + SINR_t)}{p_t} & \text{正常传输} \end{cases} \tag{16}$$

其中：W 为信道带宽；$SINR_t$ 为 t 时刻状态 s_t 对应认知用户 cr_i 的信干噪比，其计算式如下：

$$SINR_t = \frac{h_{cr_i}(c) p_{cr_i}}{n_0 + \sum_k g_k(c) p_k^{pu} + \sum_{cr_j \in K \setminus \{cr_i\}} h_{cr_j}(c) p_{cr_j}} \tag{17}$$

其中：n_0 为高斯白噪声功率；p_{cr_i} 为认知用户 cr_i 选择的功率；p_k^{pu} 为主用户 k 的发射功率；$h_{cr_i}(c)$ 为认知用户 cr_i 在信道 c 上通信时的信道增益；$g_k(c)$ 为主用户 k 在信道 c 上通信时的信道增益。

2.3　案例特征、效用值和更新方法的设计

本文按照图 3 所示的格式存储案例，选取信道增益 h_i 和 g_k 作为特征参数；存储的决策方案为进行改进学习后的最终 Q 值表；本文算法追求的目标是认知系统能量效率最大化，因此，案例效用值 E 设定为各案例达到稳态时的能量效率值。案例的更新根据效用值进行，当新案例学习得到解决方案后，与案例库中效用值最小的案例进行比较，若新案例的效用值小于最小效用值，则不更新案例库；若新案例的效用值大 于最小效用值，则完成案例更新，即用新案例替代最小效用值对应的案例存储到案例库中。

2.4　算法步骤

综上，基于案例推理和改进启发式 Q 学习的集中式认知无线电网络资源分配算法（也记为 CBR-IHAQL）具体流程如下：

a）给定 α、θ、γ、U、T_0 以及迭代次数 I，随机初始化 20 组信道增益，分别进行改进启发式 Q 学习，然后作为历史案例按格式存储到案例库中；

b）针对当前问题，根据式（2）和（3）得到匹配案例，提取其 Q 值并归一化处理，作为后续学习的初始 Q 值；

c）给定 δ_n，初始化 $G(s,a) \leftarrow 0$，$p(s,a) \leftarrow 0$，$r_{\max} \leftarrow 0$，$H(s,a)$为（0，1）间的随机数，随机选择初始状态 s_0；

d）基于当前状态 s_t，根据式（8）选择相应的动作并执行，由式（16）得到奖赏值 r_t 和下一状态 s_{t+1}；

e）根据式（4）更新 Q 值；

f）判断若 $r_t > r_{\max}$，则根据式（5）~（7）更新各动作的启发函数值、指导函数值和信息强度值，$r_{\max} \leftarrow r_t$，否则不进行更新；

g）参数更新：温度参数 T 根据下式进行更新；

$$T \leftarrow T_0 - \frac{t}{I} T_0$$

h）$s_t \leftarrow s_{t+1}$，若达到迭代次数 I，则转步骤（9）；否则，返回步骤 d）；

i）完成学习后当前问题得到解决，根据 2.3 节中设计的更新方法完成案例库的更新；

j）当有新问题到达时，转步骤 b）；否则，结束学习。

当不考虑案例更新时，算法简记为 CBR-IHAQLu。

3　算法仿真与性能分析

实验 1　改进启发式 Q 学习算法的性能

仿真中设定主用户数 M 为 3，其传输功率为 200 mW；认知用户数 K 为 6，可选的传输功率集 P={100,125,150,175,200} mW；信道数 N 为 12，其带宽 W 均为 1 MHz，且各信道增益服从均值为 1 的瑞利分布，并在学习期间内保持不变；$n_0 = 10^{-7}$ mW，T_0=0.6，θ=0.3，U=1，α=0.12，γ=0.9。本实验对 IHAQL 算法、VLRQL 算法[8]、文献[13]的 PSG-HAQL 算法以及 QL 算法[6]的性能进行对比，总迭代次数 I 为 20000 次，为了结果能更直观明了[14]，均分为 20 个学习阶段进行统计。仿真结果取 10 个新案例实验的平均值。

1）算法性能对比

图 5 和 6 分别给出了主用户以概率 λ=0.8 占用信道时，四种 Q 学习算法的认知系统能量效率和认知用户成功传输概率曲线。由图可见，随着学习时间的推移，四种 Q 学习算法的系统能量效率和认知用户成功传输概率都逐渐增加。且变化趋势一致。由于可用信道数大于用户数，四种算法趋于收敛时，所有认知用户都能以概率 1 实现成功传输，从而保证各用户都能正常通信。本文 IHAQL 算法较 VLRQL 算法、PSG-HAQL 算法和 QL 算法能够更快地达到收敛状态，且系统能量效率高。综上可见，本文算法可以快速选择最佳信道和功率，使认知用户得到更好的 QoS 保证。

chinaXiv:201810.00020v1

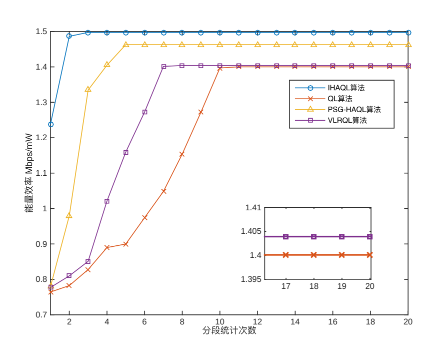


图 5　认知系统能量效率

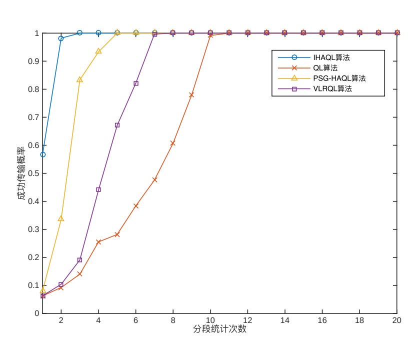


图 6　成功传输率

2）主用户数变化对系统性能的影响

图 7、8 分别给出了主用户数量变化且以概率 1 占用信道时，四种 Q 学习算法的认知系统能量效率和系统容量变化曲线。由图可见，四种 Q 学习算法的系统容量相差不大，且认知系统能量效率和系统容量都随主用户数的增加而下降，由于主用户数越大，认知用户可选择的信道越少，即受到的干扰越大，所以认知系统能量效率下降的速率越快；本文 IHAQL 算法的能量效率较 VLRQL 算法、PSG-HAQL 算法和 QL 算法的都高。

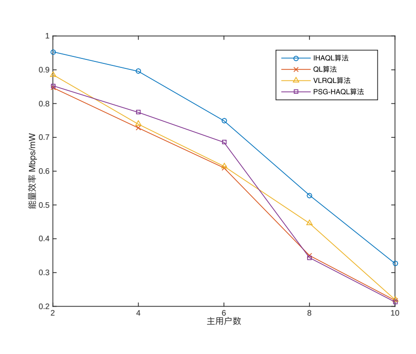


图 7 认知系统能效随主用户数变化

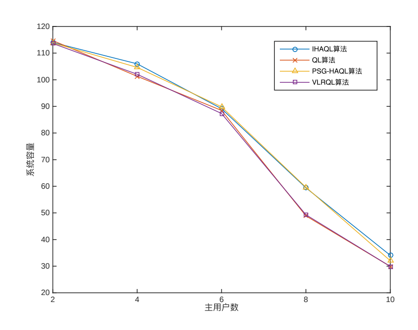


图 8 认知系统容量随主用户数变化

实验 2　基于案例推理与改进启发式 Q 学习算法（CBR-IHAQL）性能

仿真参数同实验 1。案例库中存储 20 个案例；给定 $\delta_1 = 0.8$，$\delta_2 = 0.2$，分别对应信道增益 h_i 和 g_k。总迭代次数 I 为 8000，均分为 20 个学习阶段进行统计。

1）相似值 f 对系统容量的影响

当前问题与案例库中案例的匹配相似值 f 分别为 0.1506、0.3009 和 0.5771 时，系统能量效率的变化曲线如图 9 所示。由图可见，f 越大，当前问题与存储案例的相似度越小，系统能量效率的初始值越小，收敛速度越慢，其到达稳态后的系统能效值也越小。因此，案例检索时应该选取 f 值最小的案例作为匹配案例。

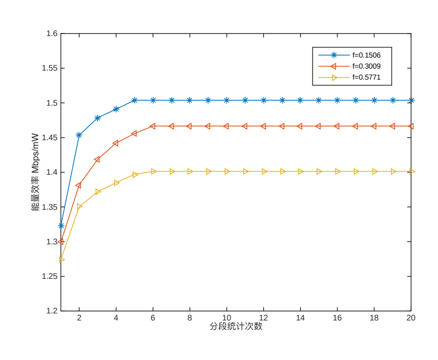


图 9 相似值 f 对系统容量的影响

2）算法性能比较

图 10 给出了 Case 分别为 20 和 40 时，CBR-IHAQL 算法、CBR-IHAQLu 算法和 IHAQL 算法的认知系统能量效率曲线，由图可见，CBR-IHAQL 算法和 CBR-IHAQLu 算法在学习初始的系统能量效率值就接近最高值，较 IHAQL 算法的收敛速度明显提高，且能达到更高的系统能效值。同时，CBR-IHAQL 算法比 CBR-IHAQLu 算法能更快收敛到更高的系统能效值。当存储的案例增加时，认知系统的能量效率增加，收敛速度也略有增加。但是当存储的案例过多时，案例匹配的搜索时间会增加，因此存储的案例数通常不会取太大，且由图可知，案例更新可以有效弥补案例不足的缺点。

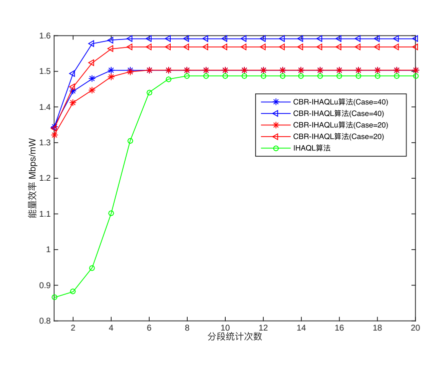


图 10 认知系统能量效率

综上可见，本文提出的基于案例推理与改进启发式 Q 学习

ChinaXiv合作期刊

算法可以快速选择最佳信道和功率，使认知用户得到更好的 QoS 保证。

4　结束语

本文主要研究了集中式认知无线电网络结构中采用案例推理与改进启发式 Q 学习的各认知用户的信道和功率分配算法。通过匹配当前问题与案例库中的案例，使 Q 学习算法在迭代初就具有接近最优解的 Q 初值，大大提高了学习效率。并在启发式 Q 学习的动作策略上引入基于信息强度的指导函数，评估各动作的重要性以指导动作的选取。仿真结果表明，该算法保证了认知用户信道和发射功率分配时的系统容量和认知用户的成功传输概率，显著提高了收敛速度。

参考文献：

[1] El Tanab M, Hamouda Walaa. Resource allocation for underlay cognitive radio networks: a survey [J]. IEEE Communications Surveys & Tutorials, 2017, 19 (2): 1249-1276.

[2] Lall S, Sadhu A K, Konar A, *et al.* Multi-agent reinfocement learning for stochastic power management in cognitive radio network [C]// Proc of International Conference on Microelectronics, Computing and Communications (MicroCom) . Piscataway, NJ: IEEE Press, 2016: 1-6.

[3] Lin Yun, Wang Chao, Wang Jiaxing, *et al.* A novel dynamic spectrum access framework based on reinforcement learning for cognitive radio sensor networks [J]. Sensors, 2016, 16 (10): 1-22.

[4] Yao Yanjun, Feng Zhiyong. Centralized channel and power allocation for cognitive radio networks: a q-learning solution [C]// Proc of Future Network & Mobile Summit. Piscataway, NJ: IEEE Press, 2010: 1-8.

[5] 伍春，江虹，易克初．聚类多 Agent 强化学习认知无线电资源分配 [J]. 北京邮电大学学报, 2014, 37 (1): 80-84. (Wu Chu, Jiang Hong, Yi Kechu. Cognitive radio resource allocation by clustering multi_agent enforcement learing [J]. Journal of Beijing University of Posts and Telecommunications, 2014, 37 (1): 80-84.)

[6] Morozs N, Clarke T, Grace D. Cognitive Spectrum management in dynamic cellular environments: a case-based Q-learning approach [J]. Engineering Applications of Artificial Intelligence, 2016, 55: 239-249.

[7] 张文柱，周雪婷，刘玉琦．认知无线电 NC-OFDM 中基于案例推理的无线资源分配 [J]. 移动通信, 2017, 41 (14): 82-88. (Zhang Wenzhu, Zhou Xueting, Liu Yuqi. Radio resource allocation based on case-reasoning in nc-ofdm systems for cognitive radio [J]. Mobile Communications, 2017, 41 (14): 82-88.)

[8] Cho B, Kim K J, Chung J W. CBR-based network performance management with multi-agent approach [J]. Cluster Computing, 2017, 20 (1): 1-11.

[9] 赖海超，赵知劲，郑仕链．应用案例推理技术的快速认知引擎 [J]. 信号处理，2012, 28 (12): 1700-1705. (Lai Haichao, Zhao Zhijin, Zheng Shilian. Fast cognitive engine using case-based reasoning [J]. Singal Processing, 2012, 28 (12): 1700-1705.)

[10] 王军红，江 虹，黄玉清，等．基于 RPkNN-Sarsa (λ) 强化学习的机器人路径规划方法 [J]. 计算机应用研究, 2013, 30 (1): 199-201. (Wang Junhong, Jiang Hong, Huang Yuqing *et al.* Method of RPkNN　Sarsa (λ) Reinforcement learning for robot path planing [J]. Application Research of Computers, 2013, 30 (1): 199-201.

[11] 冯陈伟，袁江南．基于强化学习的异构无线网络资源管理算法 [J]. 电信科学，2015, 31 (8): 99-106. (Feng Chenwei, Yuan Jiangnan. Heterogeneous wireless network resource management algorithm based on reinforcement learning [J]. Telecommunications Science, 2015, 31 (8): 99-106.

[12] Bianchi R A C, Martins M F, Ribeiro C H C, *et al.* Heuristically-accelerated multi-agent reinforcement learning [J]. IEEE Trans on Cybernetics, 2014, 44 (2): 252-265.

[13] 吴昊霖，蔡乐才，高祥．在线更新的信息强度引导启发式 Q 学习 [J]. 计算机应用研究, 2018, 35 (8): 2323-2327. (Wu Haolin, Cai Lecai, Gao Xiang. Online pheromone stringency guiding heuristically accelerated Q-learning [J]. Applications Research of Computers, 2018, 35 (8): 2323-2327.

[14] 康俊丽，郭坤祺，曹亚兰，等．一种多 agent 系统频谱接入算法 [J]. 无线通信技术, 2015, 24 (4): 7-12. (Kang Junlin, Guo Kunqi, Cao Yalan *et al.* A spectrum access algorithm for multi-agent systems [J]. Wireless Communication Technology, 2015, 24 (4): 7-12.)

chinaXiv:201810.00020v1